Large-scale analysis of the yeast proteome by multidimensional protein identification technology

Michael P. Washburn[1†], Dirk Wolters[1†], and John R. Yates III[1,2*]

We describe a largely unbiased method for rapid and large-scale proteome analysis by multidimensional liquid chromatography, tandem mass spectrometry, and database searching by the SEQUEST algorithm, named multidimensional protein identification technology (MudPIT). MudPIT was applied to the proteome of the *Saccharomyces cerevisiae* strain BJ5460 grown to mid-log phase and yielded the largest proteome analysis to date. A total of 1,484 proteins were detected and identified. Categorization of these hits demonstrated the ability of this technology to detect and identify proteins rarely seen in proteome analysis, including low-abundance proteins like transcription factors and protein kinases. Furthermore, we identified 131 proteins with three or more predicted transmembrane domains, which allowed us to map the soluble domains of many of the integral membrane proteins. MudPIT is useful for proteome analysis and may be specifically applied to integral membrane proteins to obtain detailed biochemical information on this unwieldy class of proteins.

Modern biologists can now observe quantitative changes in the expression levels of thousands of messenger RNA (mRNA) transcripts to determine the effects of a wide variety of perturbations to a cell[1]. However, there exists conflicting evidence regarding the correlation between mRNA and protein abundance levels[2–5]. Recent mathematical modeling studies have demonstrated the need to know both the mRNA and protein expression levels of genes in order to describe a gene network[6,7]. The need to complement mRNA expression analysis has resulted in the emergence of the field of proteomics to directly analyze protein expression levels from an organism.

The analysis of a proteome requires the resolution of the proteins in a sample followed by the identification of the resolved proteins. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) followed by mass spectrometry (MS) is the most widely used method of protein resolution and identification[8–10]. In 2D-PAGE, proteins are separated in one dimension by isoelectric point (pI) and in the other dimension by molecular weight (MW). High-throughput analysis of proteomes remains challenging because the individual extraction, digestion, and analysis of each spot from 2D-PAGE is a tedious and time-consuming process. As a result, the largest 2D-PAGE-based proteomic study to date identified 502 unique proteins for the *Haemophilus influenzae* proteome[11]. Portions of proteomes such as proteins with extremes in pI and molecular weight[12,13], low-abundance proteins[14–16], and membrane-associated or bound proteins[17,18] are rarely seen in a 2D-PAGE study. While efforts to alleviate the current shortcomings in 2D-PAGE continue, we are exploring non-gel-based chromatography systems to resolve and identify thousands of proteins from a biological sample[19–21].

Like 2D-PAGE, an alternative two-dimensional separation system must subject proteins or peptides to two independent separation methods and maintain the separation of two components after they have been resolved in one step[22]. A variety of efforts are underway to utilize multidimensional chromatography coupled with mass spectrometry to characterize proteomes[23]. Link *et al.* developed an online method coupling two-dimensional liquid chromatography (LC) to tandem mass spectrometry (MS/MS) (Fig. 1)[19]. In this method a pulled microcapillary column is packed with two independent chromatography phases[19]. Once a complex peptide mixture was loaded onto the system, no additional sample handling was required because the peptides eluted directly off the column and into the mass spectrometer (Fig. 1)[19]. After optimizing this system, we carried out the largest number of protein identifications in any proteome to date. By simultaneously resolving peptides and identifying their respective proteins, the system separated and identified 1,484 proteins from the *S. cerevisiae* proteome. Because the system is largely unbiased, proteins from all subcellular portions of the cell with extremes in pI, MW, abundance, and hydrophobicity were identified.

Results

The MudPIT method described is reproducible on the levels of both the chromatography and the final protein list (data not shown). Chromatographic reproducibility is described as the identification of the same peptide at the same point in the chromatography in two or more separate analyses. The results reported in this paper are from representative runs of the three separate fractions. After combining the MS/MS data generated from all three different samples, we were able to assign 5,540 peptides to MS spectra leading to the identification of 1,484 proteins from the *S. cerevisiae* proteome. A complete list of the proteins and peptides identified is available as Supplementary Table 1 in the Web Extras page of *Nature Biotechnology* Online. Each of the three preparations (soluble fraction, lightly washed insoluble fraction, and heavily washed insoluble fraction) provided unique hits to the final data set. The proteins identified in the AUTOQUEST output were further analyzed using the MIPS *S. cerevisiae* catalogs[24]. This analysis revealed that (1) our results provide a representative sampling of the yeast proteome, and (2) our MudPIT method is largely unbiased, meaning that low-abundance proteins, proteins with extremes in pI and MW, and integral membrane proteins were identified with the same sensitivity as any other protein.

[1]*Syngenta Agricultural Discovery Institute, 3115 Merryfield Row, Suite 100, San Diego, CA 92121.* [2]*Department of Cell Biology SR11, 10550 North Torrey Pines Road, The Scripps Research Institute, La Jolla, CA 92037.* *Corresponding author (jyates@scripps.edu). †These authors contributed equally to this work.

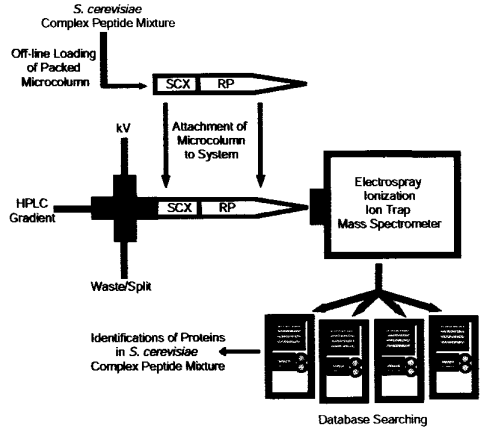

Figure 1. Multidimensional protein identification technology (MudPIT). Based on the method of Link *et al.*[19], complex peptide mixtures from different fractions of a *S. cerevisiae* whole-cell lysate were loaded separately onto a biphasic microcapillary column packed with strong cation exchange (SCX) and reverse-phase (RP) materials. After loading the complex peptide mixture into the microcapillary column, the column was inserted into the instrumental setup. Xcalibur software, HPLC, and mass spectrometer were controlled simultaneously by means of the user interface of the mass spectrometer. Peptides directly eluted into the tandem mass spectrometer because a voltage (kV) supply is directly interfaced with the microcapillary column. As described in the Experimental Protocol, peptides were first displaced from the SCX to the RP by a salt gradient and eluted off the RP into the MS/MS. In an iterative process, the microcolumn was re-equilibrated and an additional salt step of higher concentration displaced peptides from the SCX to the RP. Peptides were again eluted by an RP gradient into the MS/MS, and the process was repeated. The tandem mass spectra generated were correlated to theoretical mass spectra generated from protein or DNA databases by the SEQUEST algorithm[21].

Representative sampling of the yeast proteome. The subcellular localization catalogs from MIPS (ref. 24) allowed us to determine the similarities and differences among the three fractions (Table 1). Even though in several cases the overall numbers of proteins identified from a cellular compartment appear similar between any two samples, unique identifications were found in every sample. For example, the majority of the unique hits from the soluble fraction were proteins localized to the cytoplasm and the nuclei of *S. cerevisiae* including the transcription factor SNF5 (Codon Adaptation Index[25] (CAI) = 0.12)[26,27] and the superoxide dismutase chaperone LYS7 (CAI = 0.16)[28].

The two insoluble fractions provided greater detections and identifications of organelle proteins (Table 1). The heavily washed insoluble fraction had more hits than any other sample localized to the nucleus, mitochondria, endoplasmic reticulum, plasma membrane, and Golgi (Table 1). There were unique hits found in both the heavily washed insoluble fraction and partially washed insoluble fraction. For example, the majority of the hits to the vacuole were identified in the lightly washed insoluble fraction including the H^+-ATPase domains VMA4 (CAI = 0.27) and VMA5 (CAI = 0.24)[29].

Using the MIPS catalogs we determined that every major functional category and protein class were represented in our data (data not shown). Of the major protein classes rarely seen on 2D-PAGE, we detected and identified 32 protein kinases including the MAP kinase signal transduction pathway kinases STE7 (CAI = 0.12), STE11 (CAI = 0.15), STE20 (CAI = 0.16), and FUS3 (CAI = 0.12)[30,31]. Furthermore, we detected and identified 45 transcription factors including members of the SWI–SNF complex SNF5 (CAI = 0.12), SWI4 (CAI = 0.15), and SWI6 (CAI = 0.14)[26,27].

Of the 6,216 open reading frames in the yeast genome, 83% have CAI values between 0 and 0.20, that is, are predicted to be present at low levels (Fig. 2A). Previous proteomics studies in yeast have identified few proteins with CAIs <0.2 (refs. 4,5,32). Efforts are underway to overcome these shortcomings of 2D-PAGE, but recent evidence suggests that 2D-PAGE alone is incapable of detecting low-abundance proteins[16]. Any large-scale proteomic analysis of *S. cerevisiae* must identify proteins in this CAI range. As seen in Figure 2B, the data from our study yield a representative sample of the yeast proteome with 791 or 53.3% of the proteins identified having a CAI of <0.2. A total of 1,347 peptides were detected from the 791 proteins identified with a CAI of <0.2, an average of 1.7 peptides per protein. The number of peptides per protein increases with increasing CAI (Fig. 2C). Because CAI is considered a predictor of protein abundance[4], the most abundant proteins are the easiest to detect in any sample resulting in more peptide identifications from abundant proteins than low-abundance proteins.

Extremes of the *S. cerevisiae* proteome are well represented in our data. Because a peptide mixture is generated before the chromatography, the method should be independent of pI and MW of proteins. In two of the studies for which MW and pI were reported for the proteins identified, no protein with a MW >180 kDa or pI >10 was detected and identified[5,32]. Proteins with both acidic and basic pIs are represented in our data set. Twelve proteins with pIs <4.3 were identified, with the lowest being RPP1A (YDL081C), which has a pI of 3.82 (data not shown). Twenty-nine proteins with pIs >11 were identified, with the most basic protein identified being RPL39 (YJL189W), which has a pI of 12.55 (data not shown). In addition, proteins with MWs <10,000 and >190,000 Da are represented. For example, 24 out of 77 possible proteins with a MW in excess of 190 kDa were identified, the largest being YLR106C (CAI = 0.17) with a MW of 558,942 Da, from which four unique peptides were identified.

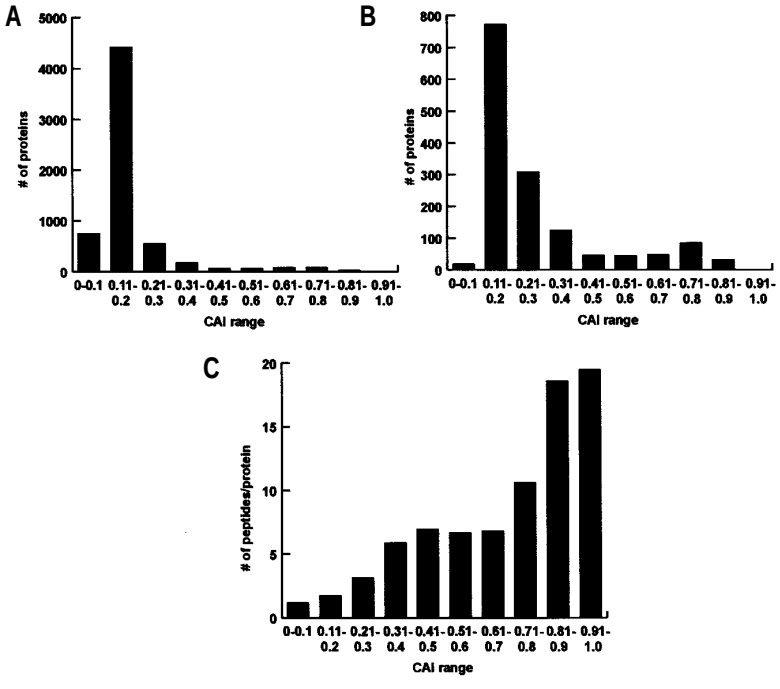

Figure 2. Codon adaptation index (CAI) distribution of the identified *S. cerevisiae* proteome and the predicted *S. cerevisiae* genome. (A) CAI distribution of the proteins predicted in the *S. cerevisiae* genome. (B) Compare this to the distribution of the proteins identified in this study over CAI ranges. In both cases, the largest protein region is found between the CAI range of 0.11 and 0.2. (C) The average number of peptides identified for each protein in a particular CAI range was determined and plotted against CAI ranges.

Table 1. Known subcellular localization of proteins identified in *S. cerevisiae* fractions[a]

Subcellular compartment	Soluble fraction[b]	Lightly washed insoluble fraction[b]	Heavily washed insoluble fraction[b]
Cell wall	2	1	1
Plasma membrane	5	18	35
Cytoplasm	286	264	274
Cytoskeleton	11	20	22
Endoplasmic reticulum	12	36	42
Golgi	3	10	16
Transport vesicles	4	14	16
Nucleus	67	122	151
Mitochondria	43	87	83
Peroxisome	2	3	3
Endosome	1	1	2
Vacuole	5	10	6
Microsomes	0	0	1
Lipid particles	0	2	3

[a]Subcellular localizations obtained from the *S. cerevisiae* subcellular localization catalog at the Munich Information Center for Protein Sequences website [24]. [b]Proteins identified in individual runs were analyzed for their subcellular localization. The subcellular localization of many of the proteins detected and identified is unknown. Therefore, not all of the proteins detected and identified are represented in this table.

Detection and identification of integral and peripheral membrane proteins. By analyzing our data set against the peripheral membrane proteins contained in the Yeast Proteome Database[33], we detected and identified 72 out of 231 possible peripheral membrane proteins. We uniquely detected 23 in the heavily washed insoluble fraction and 14 from the lightly washed insoluble fraction (data not shown).

At the MIPS website[24], the entire yeast genome has been analyzed for loci with predicted transmembrane (Tm) domains from 1 to 20 by applying the criteria of Klein *et al.*[34] and Goffeau *et al*[35]. Using these criteria, 697 proteins from the *S. cerevisiae* genome have three or more predicted Tm domains, of which we identified 131 or 19% of the total (Table 2). Of these 131 proteins, 44 were identified only in the heavily washed insoluble fraction, and 33 were identified only in the lightly washed insoluble fraction. Several of these proteins have low predicted abundances based on their CAI. For example, two unique peptides were detected for the poorly characterized protein YCR017c (CAI = 0.16), which has 15 predicted Tm domains (Table 3)[24].

The peptides detected and identified from each predicted integral membrane protein rarely covered part of or all of a predicted Tm domain (Table 3). Of the 70 peptides identified from 26 proteins with 10 or more predicted Tm domains, 4 peptides partially covered predicted Tm domains (FKS1, ALG7, and YGR125w) and 4 peptides completely covered predicted Tm domains (ALG7, ITR1, PMA1, and PMA2) (Table 3). Furthermore, 43 of the 70 peptides listed in Table 3 mapped to the largest soluble domain of the respective protein. These patterns persisted with the identifications of proteins with three to nine predicted transmembrane domains.

For example, 13 unique peptides were assigned to PMA1. PMA1 is the major isoform of the H^+-transporting P-type ATPase found in the plasma membrane[36], and a three-dimensional map of a plasma membrane H^+-ATPase from *Neurospora crassa* has been reported[37]. Of the 13 unique peptides identified from PMA1 in our analysis, 10 were from the soluble-loop domain between the fourth (amino acids 326–342) and fifth (amino acids 662–678) predicted Tm domains (Fig. 3). This gap of 342 amino acids between these two predicted Tm domains is the largest domain between two Tm domains in PMA1 and is the catalytic subunit[24]. Interestingly, the peptide of amino acids 659–680, which completely covers the fifth Tm domain, was detected and identified in our analysis (Fig. 3).

An earlier comparison of the 8 Å crystal structures of both the Ca^{2+}-ATPase from sarcoplasmic reticulum[38] and the plasma membrane H^+-ATPase from *N. crassa*[37] demonstrated that the membrane domains of both of the proteins are positioned in a highly similar fashion[39]. The crystal structure of Ca^{2+}-ATPase from sarcoplasmic reticulum, a P-type ATPase, has recently been determined[40,41]. In this crystal structure, the fifth Tm domain protrudes beyond the membrane and forms a column on which the phosphorylation domain is fixed[40,41]. We detected and identified the corresponding Tm domain in PMA1 in our analysis (Tm 5 in both P-type ATPases) (Fig. 3).

Discussion

Saccharomyces cerevisiae has been the subject of a wide variety of proteomic analyses[4,5,32,42,43], but the greatest number of proteins identified previously in a single study was 279 (ref. 32). All of these studies utilized 2D-PAGE coupled to MS, which is time-consuming as a result of the nature of spot-by-spot analysis and biased against low-abundance proteins, integral membrane proteins, and proteins with extremes in pI or MW. A substitute to 2D-PAGE/MS as the method for proteomic analyses must resolve proteins as well as 2D-PAGE, allow for the rapid identification of the proteins resolved, and deal equally with proteins, regardless of their abundance, subcellular localization, or physicochemical parameters.

To achieve the resolving power of 2D-PAGE, a multidimensional chromatography method must be used. A wide variety of systems coupling multidimensional chromatography to mass spectrometry have been described[19,23,44]. Although these methods may be suitable to automation, none identified >200 proteins from any sample. Many different types of chromatography (ion exchange, reverse phase, size exclusion) may be used in tandem so long as they are largely independent and components resolved in one dimension remain resolved in the second dimension[22]. Next, a fully automated high-throughput method is needed that combines resolution and identification removing all sample-handling steps once the sample is loaded onto the system. A fully online 2D LC/MS/MS system like MudPIT fulfills both of these requirements. Once a sample is

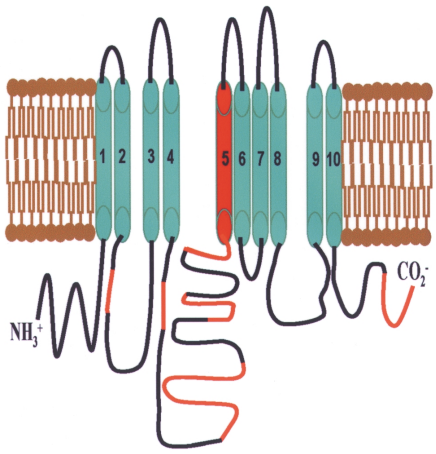

Figure 3. Peptide mapping of the integral membrane protein PMA1. A two-dimensional representation of PMA1 is displayed. Cylinders represent the predicted Tm domains as reported by MIPS (ref. 24). The protein segments between predicted Tm domains are drawn to approximate scale. Black lines and green cylinders represent segments of the protein not identified in this study. Red lines and the red cylinder represent segments of the protein identified in this study. One peptide was detected and identified between Tm domains 2 and 3, 10 peptides were detected and identified between Tm domains 4 and 5, and one peptide was detected and identified in the C terminus. We also detected and identified a peptide corresponding to Tm domain 5 in our analysis. The 320-amino acid domain between Tm domains 4 and 5 is the largest in the protein.

Table 2. Proteins identified containing three or more predicted transmembrane domains[a]

Number of predicted transmembrane domains	Number of proteins in class	Number of proteins in class identified by MudPIT	Percentage of total predicted
3	185	31	17
4	101	16	16
5	57	12	21
6	58	14	24
7	56	7	13
8	54	13	24
9	71	12	17
10	53	14	26
11	30	4	13
12	15	4	27
13	8	3	38
14	3	0	0
15	4	1	25
16	1	0	0
20	1	0	0
Totals	697	131	19

[a]The Munich Information Center for Protein Sequences website was used to obtain this information[24]. The prediction of transmembrane domains at this site is based on Klein *et al.*[34] and Goffeau *et al*[35].

loaded onto the two-dimensional column and inserted into the system (Fig. 1), no further operator interaction is needed. The major improvement over 2D-PAGE systems is that the resolution of peptides and the generation of tandem mass spectra occur simultaneously on the same sample. That is, at any given point in time, the mass spectrometer is generating tandem mass spectra to be searched against a protein database, while the HPLC and microcapillary column are resolving and eluting peptides directly into the mass spectrometer.

Although the 1,484 proteins we identified likely do not represent a complete analysis of all the proteins present in logarithmically growing cells, our method clearly provides a large-scale and global view of the *S. cerevisiae* proteome. Our methodology not only gave access to low-abundance proteins, membrane proteins, proteins with MW in excess of 180 kDa, and proteins with pIs >10, but more importantly, it did so in a largely unbiased manner. Figure 4 illustrates this point by plotting the number of proteins identified in a particular class as a percentage of the predicted proteins. The sensitivity level across the classes of proteins listed ranged from 13% of the predicted proteins identified with pIs <4.3 and MWs <10 kDa to 43% of the predicted proteins identified with pIs >11 (Fig. 4). The method has a slight bias against proteins with a pI <4.3 and MWs <10 kDa, although proteins from both of theses classes were identified. The decreased sensitivity to these classes was likely a result of a lack of tryptic peptides in the final mixture. Generally, the smaller the protein the fewer the tryptic peptides possibly generated within the mass-to-charge ratio range of the mass spectrometer. Furthermore, proteins with pIs <4.3 have fewer lysine or arginine residues that can be targeted during the endoproteinase Lys-C/trypsin digestion. Consequently, fewer peptides are generated from those acidic proteins, decreasing their chances of being identified during a MudPIT run.

The identification of integral membrane proteins by 2D-PAGE is an intensive area of research in which progress is being made[17,18]. In the most detailed proteomic analysis of a membrane from a cell to date, Molloy *et al.* identified 21 of 26 predicted integral membrane proteins from the outer membrane of *Escherichia coli* K-12 cells[45]. We identified 131 proteins with three or more predicted integral membrane proteins (Table 2) using formic acid and CNBr as the first step in the sample treatment.

Table 3. Proteins identified with 10 or more predicted transmembrane (Tm) domains[a]

Locus	Name	No. of predicted Tm domains	No. of peptides identified	Peptide hits within Tm domains[b]	Peptide hits to largest soluble domain	CAI	MW (kDa)	Membrane localization in cell
YCR017C	–	15	2	N	1	0.16	108	–
YGR032w	GSC2	13	4	N	3	0.21	217	Plasma
YIL030c	SSM4	13	1	N	0	0.17	151	–
YJL198w	–	13	1	N	1	0.18	98	–
YDR135c	YCF1	12	3	N	2	0.15	171	Vacuolar
YKL209c	STE6	12	1	N	1	0.13	145	Plasma
YLL015w	–	12	1	N	0	0.14	177	–
YLR342w	FKS1	12	6	1 P	3	0.27	215	Plasma
YGL022w	STT3	11	2	N	2	0.21	82	ER[c]
YNL268w	LYP1	11	1	N	1	0.22	68	Plasma
YNR013c	–	11	3	1 P	1	0.19	99	Plasma
YPL058c	PDR12	11	6	N	3	0.29	171	–
YBR068c	BAP2	10	2	N	0	0.16	68	Plasma
YBR243c	ALG7	10	2	1 P, 1 C	0	0.13	50	ER
YDR342c	HXT7	10	1	N	1	0.52	63	Plasma
YDR343c	HXT6	10	2	N	1	0.52	63	Plasma
YDR345c	HXT3	10	1	N	1	0.49	63	Plasma
YDR497c	ITR1	10	1	C	0	0.19	64	Plasma
YER119c	–	10	1	P	0	0.10	49	–
YFL025c	BST1	10	1	N	0	0.13	118	ER
YGL008c	PMA1	10	13	1 C	10	0.73	100	Plasma
YGR125w	–	10	1	1 P	0	0.12	117	–
YHR094c	HXT1	10	1	N	1	0.41	63	Plasma
YLL061w	MMP1	10	1	N	0	0.13	64	–
YOR328w	PDR10	10	1	N	1	0.13	176	Plasma
YPL036w	PMA2	10	11	1 C	10	0.30	102	Plasma

[a]The Munich Information Center for Protein Sequences website was used to obtain this information. The prediction of transmembrane domains at this site is based on Klein *et al.*[34] and Goffeau *et al.*[35] [b]Abbreviations: N, none; P, partially covers a transmembrane domain; C, completely covers a transmembrane domain. [c]Endoplasmic reticulum.

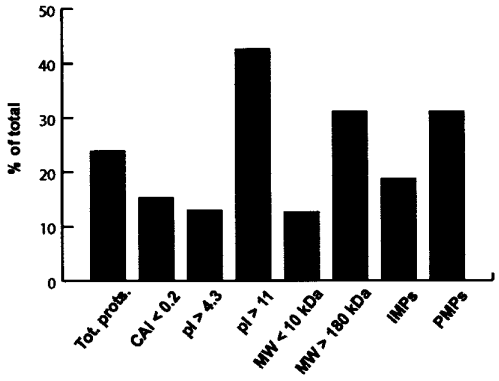

Figure 4. Sensitivity of MudPIT to a wide variety of protein classes. The percentage of proteins identified in this study from a variety of protein classes is presented. The percentages were determined by dividing the number of proteins identified in the study in each category shown by the total number of predicted proteins from each category shown. MIPS (ref. 24) and the Yeast Proteome Database[33] were used to obtain the predicted numbers of proteins from *S. cerevisiae* in each class. From left to right are the percentages identified of total proteins, proteins with a CAI <0.2, proteins with a pI < 4.3, proteins with a pI >11, proteins with a MW <10kDa, proteins with a MW >180 kDa, integral membrane proteins (IMPs) with three or more predicted transmembrane domains, and peripheral membrane proteins (PMPs).

Because formic acid is an organic acid, it partially solubilized the membrane portions of the cell in our heavily and lightly washed insoluble fractions. Then, CNBr cleaved off the soluble portions of the integral membrane proteins as large domains that were subjected to additional proteolysis. Peptides detected and identified from integral membrane proteins rarely contained any portion of a predicted transmembrane domain (Table 3). When multiple hits were obtained to a particular integral membrane protein, the peptides identified typically localized to the largest soluble loop between two predicted transmembrane domains in the protein (Table 3 and Fig. 3). In the instance of PMA1, we identified a Tm domain that may have unique functional significance (Fig. 3). On the basis of the crystal structure of another P-type ATPase (refs 40,41) and the similarities of P-type ATPases (ref. 39), Tm domain 5 in PMA1 may protrude beyond the plasma membrane and provide a column on which the catalytic domain rests. Based on the results, our method may be useful for localizing predicted integral membrane proteins to particular membranes in a cell and for providing support for predicted folding of proteins within the membrane.

Proteomics is beginning to develop the methodology needed for comprehensive high-throughput quantitative analyses of proteomes. The method described in this work is a major step toward comprehensive high-throughput methods, because not only were low-abundance proteins detected and identified, but peripheral and integral membrane proteins were also detected. MudPIT alone is not particularly quantitative. In general, the more abundant a protein, the more peptides identified from a protein. Only when emerging quantitative proteomic methods[46–49] are combined with MudPIT will true large-scale analysis of protein expression changes be possible. The combination of MudPIT with quantitative methods will allow for the integration of mRNA and protein expression levels needed to fully understand gene networks[6,7].

Experimental protocol

Materials. Standard laboratory chemicals used in this work and acid-washed glass beads were obtained from Sigma (St. Louis, MO). Sodium vanadate ($NaVO_3$), sodium fluoride (NaF), sodium pyrophosphate ($Na_4P_2O_7$), formic acid, and cyanogen bromide (CNBr) came from Aldrich (Milwaukee, WI). Poroszyme bulk immobilized trypsin was a product of Applied Biosystems (Framingham, MA). HPLC-grade acetonitrile (ACN) and HPLC-grade methanol were purchased from Fischer Scientific (Fair Lawn, NJ). Endoproteinase Lys-C was purchased from Roche Diagnostics (Indianapolis, IN). Difco Dextrose, tryptone, and yeast extract were products of BD Biosciences (Sparks, MD). Heptafluorobutyric acid (HFBA) was obtained from Pierce (Rockford, IL). Glacial acetic acid was purchased from Malinckrodt Baker Inc. (Paris, KY).

Growth and lysis of *S. cerevisiae*. Strain BJ5460 (ref. 50) was grown to mid-log phase (OD 0.6) in YPD at 30°C. To generate three fractions to analyze, two separate groups of cells were treated in the following manner. Cells were solubilized in lysis buffer (310 mM NaF, 3.45 mM $NaVO_3$, 50 mM Tris, 12 mM EDTA, 250 mM NaCl, 140 mM dibasic sodium phosphate pH 7.60) and disrupted in the presence of glass beads in a Mini-BeadBeater (BioSpec Products, Bartlesville, OK) as described[19]. After removal of the supernatants, the remaining two pellets were subjected to additional washing as follows. Each pellet was washed by adding 1× PBS (1.4 mM NaCl, 0.27 mM KCl, 1 mM Na_2HPO_4, 0.18 mM dibasic potassium phosphate, pH 7.4) to the tube, vortexed for 2 min, and pelleted by centrifugation at 14,000 r.p.m. for 10 min in the Eppendorf microfuge. One pellet (to be named the lightly washed insoluble pellet) was washed once in this fashion followed by lyophilization to dryness in a Speed Vac SC 110 (Savant Instruments, Holbrook, NY). The second pellet (to be named the heavily washed insoluble pellet) was washed 3× in this way, followed by lyophilization to dryness.

Digestion of soluble fraction. After adjusting the pH to 8.5 with 1 M ammonium bicarbonate (AmBic), the protein concentration was determined by the Bradford assay. The sample was sequentially solubilized in 8 M urea, reduced by adding dithiothreitol to 1 mM, and carboxyamidomethylated in 10 mM iodoacetamide. After digestion with Endoproteinase Lys-C as described[19], the solution was diluted to 2 M urea with 100 mM AmBic, pH 8.5 followed by the addition of $CaCl_2$ to 1 mM. Finally, 3 µl of Porosyzme immobilized trypsin were added and incubated overnight at 37°C while rotating. After removal of the Porosyzme immobilized trypsin beads by centrifugation, a solid-phase extraction with SPEC-PLUS PTC18 cartridges (Ansys Diagnostics, Lake Forest, CA) was carried out on the supernatant according to the manufacturer's instructions to concentrate the complex peptide mixtures and buffer exchange the mixtures into 5% ACN, 0.5% acetic acid. Samples not immediately analyzed were stored at –80°C. After the preparation of the complex peptide mixture, amino acid analysis (Macromolecular Structure Facility, Department of Biochemistry, Michigan State University) was carried out on each sample.

Digestion of insoluble fractions. The lyophilized heavily washed and lightly washed insoluble fractions were treated separately by adding 100 µl of 90% formic acid and incubating for 5 min at room temperature. After adding 100 mg of CNBr, the samples were incubated overnight at room temperature in the dark. On the following day, the pH was adjusted to 8.5 by the addition of MilliQ H_2O and solid AmBic. Each fraction was lyophilized to ~200 µl. From this point forward, the samples were treated identically to the soluble fraction.

Multidimensional protein identification technology (MudPIT). Each sample was subjected to MudPIT analysis with modifications to the method described by Link *et al*[19]. A quaternary Hewlett-Packard 1100 series HPLC was directly coupled to a Finnigan LCQ ion trap mass spectrometer equipped with a nano-LC electrospray ionization source[51]. A fused-silica microcapillary column (100 µm i.d. × 365 µm o.d.) was pulled with a Model P-2000 laser puller (Sutter Instrument Co., Novato, CA) as described[51]. The microcolumn was first packed with 10 cm of 5 µm C_{18} reverse-phase material (XDB-C18, Hewlett-Packard) followed by 4 cm of 5 µm strong cation exchange material (Partisphere SCX; Whatman, Clifton, NJ). Approximately 420 µg of the soluble fraction, 440 µg of the lightly washed insoluble fraction, and 490 µg of the heavily washed insoluble fraction were loaded onto three separate microcolumns for the analysis of each fraction. After loading the microcapillary column, the column was placed in-line with the system (Fig. 1) as described[19]. A fully automated 15-step chromatography run was carried out on each sample. The four buffer solutions used for the chromatography were 5% ACN/0.02% HFBA (buffer A), 80% ACN/0.02% HFBA (buffer B), 250 mM ammonium acetate/5% ACN/0.02% HFBA (buffer C), and 500 mM ammonium acetate/5% ACN/0.02% HFBA (buffer D). The first step of 80 min consisted of a 70 min gradient from 0 to 80% buffer B and a 10 min hold at 80% buffer B. The next 12 steps were 110 min each with the following profile: 5 min of 100% buffer A, 2 min of *x*% buffer C, 3 min of 100% buffer A, a 10 min gradient from 0 to 10% buffer B, and a 90 min gradient from 10 to 45% buffer B. The 2 min buffer C percentages (*x*) in steps 2–13 were as follows: 10, 20, 30, 40, 50, 60, 70, 80, 90, 90, 100, and 100%. Step 14

consisted of the following profile: a 5 min 100% buffer A wash, a 20 min 100% buffer C wash, a 5 min 100% buffer A wash, a 10 min gradient from 0 to 10% buffer B, and a 90 min gradient from 10 to 45% buffer B. Step 15 was identical to step 14 except that the 20 min salt wash was with 100% buffer D.

SEQUEST analysis and AUTOQUEST output. The SEQUEST algorithm[21] was run on each of the three data sets against the yeast_orfs.fasta database from the National Center for Biotechnology Information. The AUTOQUEST software package displayed the output, listing protein loci with the number of peptides assigned to each locus. Because CNBr cleaves at methionine residues and leaves either homoserine (Hse) or Hse lactone[52], the MS/MS data resulting from the two samples treated with CNBr/formic acid had to be independently analyzed twice with SEQUEST[21]. For each run, the differential search modification was engaged and set to either –30 for Hse or –48 for Hse lactone. We used conservative criteria to determine the protein content of our samples based on those described[19]. Peptides identified by SEQUEST may have three different charge states (+1, +2, or +3), each of which results in a unique spectrum for the same peptide. Except in rare instances, an accepted SEQUEST result had to have a ΔCn score of at least 0.1 (regardless of charge state[21]). Peptides with a +1 charge state were accepted if they were fully tryptic and had a cross correlation (Xcorr) of at least 1.9. Peptides with a +2 charge state were accepted if they were fully tryptic or partially tryptic between the Xcorr ranges of 2.2 and 3.0. Partially tryptic peptides were especially relevant in the two samples where CNBr was used. Peptides with a +2 charge state with an Xcorr >3.0 were accepted regardless of their tryptic nature. Finally, +3 peptides were only accepted if they were fully or partially tryptic and had an Xcorr >3.75. We manually confirmed each SEQUEST result from every protein identified by four or fewer peptides using criteria described[19]. When five or more peptides were identified from a protein we manually confirmed that at least one of the SEQUEST results fit criteria described[19].

Note: Supplementary information can be found on the Nature Biotechnology *website in Web Extras (http://biotech.nature.com/web_extras).*

Acknowledgments

The authors thank Jimmy Eng, David Schieltz, David Tabb, and Laurence Florens for valuable discussions during the preparation of this manuscript. The authors acknowledge funding from the National Institutes of Health R33CA81665-01 and RR11823-03. M.P.W. acknowledges support from genome training grant T32HG000035-05. Saccharomyces cerevisiae strain BJ5460 was a generous gift from Steve Hahn of the Fred Hutchinson Cancer Research Center (Seattle, WA).

Received 21 August 2000; accepted 30 November 2000

1. Lockhart, D.J. & Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
2. Kawamoto, S., Matsumoto, Y., Mizuno, K., Okubo, K. & Matsubara, K. Expression profiles of active genes in human and mouse livers. *Gene* **174**, 151–158 (1996).
3. Anderson, L. & Seilhamer, J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18**, 533–537 (1997).
4. Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S. & Garrels, J.I. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368 (1999).
5. Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
6. Hatizmanikatis, V. & Lee, K.H. Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metabol. Eng.* **1**, 275–281 (1999).
7. Hatzimanikatis, V., Choe, L.H. & Lee, K.H. Proteomics: theoretical and experimental considerations. *Biotechnol. Prog.* **15**, 312–318 (1999).
8. Hanash, S.M. Biomedical applications of two-dimensional electrophoresis using immobilized pH gradients: current status. *Electrophoresis* **21**, 1202–1209 (2000).
9. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
10. Washburn, M.P. & Yates, J.R. Analysis of the microbial proteome. *Curr. Opin. Microbiol.* **3**, 292–297 (2000).
11. Langen, H. *et al.* Two-dimensional map of the proteome of *Haemophilus influenzae*. *Electrophoresis* **21**, 411–429 (2000).
12. Oh-Ishi, M., Satoh, M. & Maeda, T. Preparative two-dimensional gel electrophoresis with agarose gels in the first dimension for high molecular mass proteins. *Electrophoresis* **21**, 1653–1669 (2000).
13. Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F. & Sanchez, J.C. The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis* **21**, 1104–1115 (2000).
14. Fountoulakis, M., Takacs, M.F., Berndt, P., Langen, H. & Takacs, B. Enrichment of low abundance proteins of *Escherichia coli* by hydroxyapatite chromatography. *Electrophoresis* **20**, 2181–2195 (1999).
15. Fountoulakis, M., Takacs, M.F. & Takacs, B. Enrichment of low-copy-number gene products by hydrophobic interaction chromatography. *J. Chromatogr. A* **833**, 157–168 (1999).
16. Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional electrophoresis-based proteome analysis. *Proc. Natl. Acad. Sci. USA* **97**, 9390–9395 (2000).
17. Molloy, M.P. Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Anal. Biochem.* **280**, 1–10 (2000).
18. Santoni, V., Molloy, M. & Rabilloud, T. Membrane proteins and proteomics: un amour impossible? *Electrophoresis* **21**, 1054–70 (2000).
19. Link, A.J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999).
20. McCormack, A.L. *et al.* Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69**, 767–776 (1997).
21. Eng, J.K., McCormack, A.L. & Yates, J.R.I. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
22. Giddings, J.C. Concepts and comparisons in multidimensional chromatography. *J. High Res. Chromatogr.* **10**, 319–323 (1987).
23. Washburn, M.P. & Yates, J.R. Novel methods of proteome analysis: multidimensional chromatography and mass spectrometry. *Proteomics: A Current Trends Supplement*, 28–32 (2000).
24. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
25. Sharp, P.M. & Li, W.H. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
26. Peterson, C.L. & Workman, J.L. Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Curr. Opin. Genet. Dev.* **10**, 187–192 (2000).
27. Cairns, B.R., Kim, Y.J., Sayre, M.H., Laurent, B.C. & Kornberg, R.D. A multisubunit complex containing the SWI1/ADR6, SWI2/SNF2, SWI3, SNF5, and SNF6 gene products isolated from yeast. *Proc. Natl. Acad. Sci. USA* **91**, 1950–1954 (1994).
28. Culotta, V.C. *et al.* The copper chaperone for superoxide dismutase. *J. Biol. Chem.* **272**, 23469–23472 (1997).
29. Liu, Q. *et al.* Site-directed mutagenesis of the yeast V-ATPase A subunit. *J. Biol. Chem.* **272**, 11750–11756 (1997).
30. Lee, B.N. & Elion, E.A. The MAPKKK Ste11 regulates vegetative growth through a kinase cascade of shared signaling components. *Proc. Natl. Acad. Sci. USA* **96**, 12679–12684 (1999).
31. Sprague, G.F., Jr. Control of MAP kinase signaling specificity or how not to go HOG wild. *Genes Dev.* **12**, 2817–2820 (1998).
32. Perrot, M. *et al.* Two-dimensional gel protein database of *Saccharomyces cerevisiae* (update 1999). *Electrophoresis* **20**, 2280–2298 (1999).
33. Costanzo, M.C. *et al.* The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* **28**, 73–76 (2000).
34. Klein, P., Kanehisa, M. & DeLisi, C. The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta* **815**, 468–476 (1985).
35. Goffeau, A., Nakai, K., Slonimski, P., Risler, J.L. & Slominski, P. The membrane proteins encoded by yeast chromosome III genes. *FEBS Lett.* **325**, 112–117 (1993).
36. Ambesi, A., Miranda, M., Petrov, V.V. & Slayman, C.W. Biogenesis and function of the yeast plasma-membrane H(+)-ATPase. *J. Exp. Biol.* **203**, 155–160 (2000).
37. Auer, M., Scarborough, G.A. & Kuhlbrandt, W. Three-dimensional map of the plasma membrane H+-ATPase in the open conformation. *Nature* **392**, 840–843 (1998).
38. Zhang, P., Toyoshima, C., Yonekura, K., Green, N.M. & Stokes, D.L. Structure of the calcium pump from sarcoplasmic reticulum at 8-Å resolution. *Nature* **392**, 835–839 (1998).
39. Kuhlbrandt, W., Auer, M. & Scarborough, G.A. Structure of the P-type ATPases. *Curr. Opin. Struct. Biol.* **8**, 510–516 (1998).
40. McIntosh, D.B. Portrait of a P-type pump. *Nat. Struct. Biol.* **7**, 532–535 (2000).
41. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature* **405**, 647–655 (2000).
42. Shevchenko, A. *et al.* Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93**, 14440–14445 (1996).
43. Garrels, J.I. *et al.* Proteome studies of *Saccharomyces cerevisiae*: identification and characterization of abundant proteins. *Electrophoresis* **18**, 1347–1360 (1997).
44. Nilsson, C.L. & Davidsson P. New separation tools for comprehensive studies of protein expression by mass spectrometry. *Mass Spectrom. Rev.* **19**, 390–397 (2000).
45. Molloy, M.P. *et al.* Proteomic analysis of the *Escherichia coli* outer membrane. *Eur. J. Biochem.* **267**, 2871–2881 (2000).
46. Pasa-Tolic, L. *et al.* High throughput proteome-wide precision measurements of protein expression using mass spectrometry. *J. Am. Chem. Soc.* **121**, 7949–7950 (1999).
47. Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96**, 6591–6596 (1999).
48. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
49. Münchbach, M., Quadroni, M., Miotto, G. & James, P. Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal. Chem.* **72**, 4047–4057 (2000).
50. Jones, E.W. Tackling the protease problem in *Saccharomyces cerevisiae*. *Methods Enzymol.* **194**, 428–453 (1991).
51. Gatlin, C.L., Kleemann, G.R., Hays, L.G., Link, A.J. & Yates, J.R. Protein identification at the low femtomole level from silver-stained gels using a new fritless electrospray interface for liquid chromatography-microspray and nanospray mass spectrometry. *Anal. Biochem.* **263**, 93–101 (1998).
52. Aitken, A., Geisow, M.J., Findlay, J.B.C., Holmes, C. & Yarwood, A. Peptide preparation and characterization. In *Protein sequencing: a practical approach* (eds Findlay, J.B.C. & Geisow, M.J.) 43–68 (IRL Press, New York, NY; 1989).